Published online August 17, 2005

# Comparisons of Two Statistical Models for Evaluating Boll Retention in Cotton

Jixiang Wu, Johnie N. Jenkins,* Jack C. McCarty, Jr., and Clarence E. Watson

## ABSTRACT

**Boll number is one of the most important traits related to yield of upland cotton (*Gossypium hirsutum* L.). Evaluation of boll retention properties at different fruiting sites would provide useful information for cotton breeding and cotton growth management. The presence or absence of a boll at each fruiting position can be considered as binomially distributed. In this study, 188 upland cotton recombinant inbred (RI) lines, two parental lines, and a control cultivar, Stoneville 474, were used. These lines were planted at Mississippi State, MS in 1999. The data set was analyzed by the mixed linear model and logistic regression model. The results showed that the boll retention for the first position was significantly different among nodes but expressed similar total numbers from the first position among RI lines. Estimates for boll retention were similar for both models; however, the logistic regression model gave smaller confidence intervals for each estimate than the mixed linear model.**

Boll number is one of the most important traits related to yield of upland cotton (*Gossypium hirsutum* L.). The transformation and development of bolls on a plant is time and space dependent. Some researchers have focused on studying boll retention properties at different nodes and positions as well as total boll number per plant (Jenkins et al., 1990a, 1990b; Kerby et al., 1987; Jenkins and McCarty, 1995; Shoemaker, 2000). This research not only evaluated the positional contributions to total yield production but also evaluated growth behavior such as earliness. Previous research showed that bolls from first position contribute 66–75%, and bolls from second position 18–21%, to total yield of modern cultivars (Jenkins et al., 1990a, 1990b; Jenkins and McCarty, 1995; Kerby et al., 1987).

Generally, this space-dependent character, boll number or boll retention, is treated as a continuously distributed variable, which can be analyzed by the analysis of variance (ANOVA) models; however, one potential problem is that the error variation of boll retention could be quite different across positions and nodes due to changes in environmental and physiological conditions during the growing season. A data set with heterogeneity of random error variations would violate one of the requirements of the ANOVA model. As previously stated, boll retention at different fruiting sites could be correlated and have different error variations; thus the mixed model with different error variance–covariance structures may be used to improve the statistical testing powers (Littell et al., 1996).

On the other hand, a boll at a specific fruiting site on a plant is expressed as either present or absent, and thus it can be considered as binomially distributed. Boll retention varies among different fruiting sites. For example, boll retention at the first position in the middle of the plant is generally greater than that at other positions. Logistic regression analysis is often used to investigate the relationship between a binary trait and a set of explanatory variables. Several books have discussed logistic regression (Collett, 1991; Agresti, 1990, 1996; Cox and Snell, 1989; Hosmer and Lemeshow, 1996; McCullagh and Nelder, 1989). Currently, statistical software packages such as SAS are available for logistic regression analysis. In addition, SAS version 8.0 enables researchers to specify categorical variables as explanatory variables in the model (SAS Inst., 1999b).

In the current study, data from first-position bolls of 188 upland cotton RI lines, two parental lines, and one commercial cultivar in 1 yr, were used. Data analyses were conducted subject to the mixed linear model with four error structures and subject to the logistic regression model. Boll retention and its 95% confidence intervals at different fruiting nodes were calculated for both the mixed linear model and the logistic model. The purposes of this study were to compare the estimates of boll retention and their statistical precision. The results should help researchers determine which model should be used to analyze a binary-type trait.

## MATERIALS AND METHODS

### Materials

One hundred eighty-eight RI lines ($F_8$) were developed by single-hill (bulked progeny row) procedure (Fehr, 1987) from the *G. hirsutum* intraspecific cross HS46 ($P_1$) × MARCABUCAG8US-1-88 ($P_2$) (Shappley et al., 1998a, 1998b). A cross between $P_1$ and $P_2$ was made at Mississippi State, MS in 1991, and the $F_1$ generation was grown in 1992. One hundred $F_2$ seeds from one $F_1$ individual were planted in the greenhouse and selfed in 1992. The $F_3$ seeds were planted in 12-m single-row plot (named as single hill) at Mississippi State in the spring of 1994, and plants were self-pollinated and bulked by progeny row. In the winter of 1994, $F_4$ selfed seeds were sent to a nursery in Mexico for generation increase by selfing and bulked by progeny row to obtain $F_5$ seeds. In the spring of 1995, two-row $F_5$ plots from each $F_2$–derived family were planted, and 25 individual plants were selfed and harvested to obtain $F_6$ seeds. In the winter of 1996, one seed from each of 25 selfed plants from each $F_2$–derived family was sent to Mexico. Up to eight plants from each family were selfed to produce $F_7$ seeds.

J. Wu, Dep. of Plant and Soil Sci., Mississippi State Univ., Mississippi State, MS 39762; J.N. Jenkins and J.C. McCarty, Jr., Crop Sci. Res. Lab., USDA-ARS, Mississippi State, MS 39762; and C.E. Watson, MAFES Administration, Mississippi State Univ., Mississippi State, MS 39762. Contribution of the USDA-ARS in cooperation with the Mississippi Agric. and Forestry Exp. Stn. Received 13 Aug. 2004. *Corresponding author (jnjenkins@msa-msstate.ars.usda.gov).

Published in Agron. J. 97:1291–1294 (2005).
Cotton
doi:10.2134/agronj2004.0216
© American Society of Agronomy
677 S. Segoe Rd., Madison, WI 53711 USA

**Abbreviations:** CIL, confidence interval length; RI, recombinant inbred (lines).

**Table 1. Different variance–covariance structures used for mixed linear model analyses.**

| Structure | Description | Parameters | $(i,j)^{th}$ element |
|---|---|---|---|
| CSH | Heterogeneous compound symmetry | 19 | $\sigma_i\,\sigma_j[\rho 1(i \neq j) + 1(i = j)]$ |
| ARH(1) | Heterogeneous autoregressive (1) | 19 | $\sigma_i\,\sigma_j\rho^{\lvert i-j \rvert}$ |
| UN | Unstructured | 18 × 19/2 | $\sigma_{ij}$ |
| CS† | Compound symmetry | 2 | $\sigma_G^2 + \sigma^2(i = j)$ |

† CS is a general linear model (GLM) in which all nodes were assumed to have the same residual variance. In the other three structures, all nodes were assumed to have different residual variances. For detailed information, see SAS OnLine Doc Version 8 (SAS Inst., 1999b).

In the winter of 1998, up to eight individual plant progenies from each of 94 $F_2$–derived families were planted and hand-harvested separately ($F_8$ seeds). Two lines were randomly chosen from each $F_2$–derived family to reduce the population size.

The 188 RI lines (two lines from each of 94 $F_2$–derived families), two parental lines, and one commercial cultivar, Stoneville 474, were grown at the Plant Science Research Center, Mississippi State, MS in 1999 with four replications. Plots consisted of two rows, 12 m long with row spacing of 0.97 m. The field soil was a leeper silty clay loam. Before boll sample hand picking and machine harvest, 10 normal plants (no aborted terminals) in each two-row plot were randomly selected to determine boll retention for the first position from Main-Stem Nodes 5 to 22. Data were recorded as boll present = 1 or absent = 0. For each plot, boll retention of the first position for each node was calculated as number of bolls present divided by number of plants.

## Methods

The linear model used for mixed linear approach is $y_{ij} = \mu + N_i + e_{ij}$, where $\mu$ is the population mean, $N_i$ is the node effect, and $e_{ij}$ is the residual. The $e_{ij}$ could have some genetic correlation at different nodes for the same genotype and different residual variances; thus, we consider four types of variance–covariance structures [CSH, ARH(1), UN, and CS; see Table 1 for the definitions] in the linear model (Littell et al., 1996; SAS Inst., 1999b), among them, CS type is equivalent to the linear model $y_{ij} = \mu + N_i + G_j + e_{ij}$, which can be also analyzed by GLM or ANOVA method, and $G_j$ is genotypic effect and is considered a random effect. The least-squared means and standard errors for boll retention at each node were estimated for four different structures. Confidence interval length (CIL) of 95% for each parameter was calculated based on each standard error. The formula used for the calculation of CIL of 95% is: CIL = 2 × $t_{0.025}$ × SE, where SE is the standard error for boll retention at a specific node.

**Table 2. Statistical properties for boll retention for different nodes of recombinant inbred lines.**

| Node | Boll retention | CR† | CCR‡ |
|---|---|---|---|
| | % | | |
| 5 | 20.14 | 4.74 | 4.74 |
| 6 | 37.58 | 8.85 | 13.60 |
| 7 | 43.40 | 10.22 | 23.82 |
| 8 | 41.98 | 9.89 | 33.71 |
| 9 | 39.55 | 9.32 | 43.03 |
| 10 | 37.77 | 8.90 | 51.92 |
| 11 | 37.60 | 8.86 | 60.78 |
| 12 | 33.60 | 7.91 | 68.70 |
| 13 | 31.40 | 7.40 | 76.09 |
| 14 | 27.63 | 6.51 | 82.60 |
| 15 | 24.74 | 5.83 | 88.43 |
| 16 | 21.28 | 5.01 | 93.44 |
| 17 | 13.55 | 3.19 | 96.63 |
| 18 | 8.13 | 1.92 | 98.55 |
| 19 | 3.65 | 0.86 | 99.41 |
| 20 | 1.81 | 0.43 | 99.83 |
| 21 | 0.46 | 0.11 | 99.94 |
| 22 | 0.24 | 0.06 | 100.00 |

† CR = contribution rate.
‡ CCR = cumulated contribution rate.

Total number of bolls present for each node and each genotype over replications was used for logistic regression model. In the logistic regression analysis, the link function of logit was employed. The model used for logistic regression was $\pi(\eta_{ij}) = \pi(\mu, N, G_j) = e^{\mu+N_i+G_j}/(1 + e^{\mu+N_i+G_j})$, where the definitions $\mu$, $N_i$, and $G_j$ have been stated above. $\hat{\eta}_{ij} = \hat{\mu} + \hat{N}_i + \hat{G}_j$ and standard error $\hat{\sigma}(\eta_{ij})$ were calculated. The boll retention was estimated by $\hat{\pi}(\hat{\eta}_{ij})= e^{\hat{\eta}_{ij}}/(1 + e^{\hat{\eta}_{ij}})$ and 95% CIL was estimated via $\hat{\pi}[\hat{\eta}_{ij} + z_{0.025} * \hat{\sigma}(\hat{\eta}_{ij})] - \hat{\pi}[\hat{\eta}_{ij} - z_{0.025} * \hat{\sigma}(\hat{\eta}_{ij})]$ for the logistic regression model (SAS Inst., 1999a). All data analyses were conducted using SAS 8.0 (SAS Inst., 1999b).

# RESULTS

## Phenotypic Data

Mean boll retention over all RI cotton lines for different nodes is summarized in Table 2. No fruiting node (Position 1) had boll retention greater than 50%. Node 5 had approximately 20% boll retention while Nodes 6 through 13 had greater than 30% boll retention. Only Nodes 7 and 8 had greater than 40% boll retention. Node 7 had the highest boll retention, reaching approximately 44%. Above Node 7, boll retention decreased. Normally, bolls from the middle nodes account for the majority of the contribution to total boll number per plant. For example, the contribution from Nodes 6 through 15 accounted for 84% of total boll number for the first position. Contribution from Nodes 7–12 accounted for 55%.

## Mixed Linear Model Analysis

Sum of squares for boll retention based on genotype × node means at the first position obtained using the ANOVA approach are summarized in Table 3. Both genotype and node had significant impacts on boll retention at the first position. To further determine the relative importance of genotypic and node effects contributing to the phenotypic variance, we considered both genotypic and node effects as random, and variance components for genotypic and node effects were estimated using the results listed in Table 3. Node effects contributed to 78.6% of total variation while genotypic

**Table 3. Sum of squares for boll retention (%) among genotypes and nodes by the general linear model.**

| Source | df† | SS‡ | *F* value |
|---|---|---|---|
| Genotype | 190 | 24 000 | 2.2** |
| Node | 17 | 777 000 | 770.8** |
| Residual | 3230 | 191 000 | |

** Significant at probability level of 0.01.
† df = degrees of freedom.
‡ SS = sum of squares.

**Table 4. Estimates of residual variance for three variance–covariance structures using the mixed linear model approach.†**

| Node | CSH‡ | ARH(1)§ | UN¶ |
|---|---|---|---|
| 5 | 108.00 | 97.02 | 98.36 |
| 6 | 128.20 | 121.60 | 123.20 |
| 7 | 93.99 | 91.63 | 92.91 |
| 8 | 103.60 | 102.30 | 102.90 |
| 9 | 102.50 | 108.50 | 99.72 |
| 10 | 81.15 | 94.89 | 80.89 |
| 11 | 84.78 | 93.28 | 81.35 |
| 12 | 71.23 | 85.45 | 70.89 |
| 13 | 69.01 | 85.36 | 68.67 |
| 14 | 64.35 | 75.32 | 65.21 |
| 15 | 60.10 | 64.11 | 62.04 |
| 16 | 73.99 | 68.63 | 76.20 |
| 17 | 54.78 | 47.66 | 57.16 |
| 18 | 37.12 | 31.32 | 37.73 |
| 19 | 10.81 | 9.80 | 11.15 |
| 20 | 5.30 | 5.13 | 5.53 |
| 21 | 1.38 | 1.32 | 1.40 |
| 22 | 0.60 | 0.58 | 0.60 |

† Note: all variance estimates were significant at $\alpha < 0.001$.
‡ CSH = heterogeneous compound symmetry.
§ ARH(1) = heterogeneous autoregressive(1).
¶ UN = unstructured.

effects contributed to 1.7% of total variation. Thus, the data suggested that node effects had more important impact on boll retention than genotypic effects in this study, which was almost negligible. The residual including node × genotype interaction effects contributed to 19.7% of total variation in boll retention.

Residual variances for boll retention on different nodes estimated by the mixed linear model approach for three repeated measurement variance–covariance structures [CSH, ARH(1), and UN] are summarized in Table 4. The residual variances obtained by these three variance–covariance structures varied among nodes, and they were similar for these three variance–covariance structures.

## Logistic Regression Analysis

Both genotype and node were considered as categorical explanatory variables in the logistic regression analysis. A stepwise selection procedure was applied to choose the candidate explanatory variables during the analysis. In this study, only node was selected to have significant effects on boll retention in the logistic regression model (Table 5). Nodes 5 through 16 expressed positive effects while Node 17 and above expressed negative effects, indicating that Nodes 5 through 16 had higher boll retention than Node 17 and above. Therefore, boll retention could be estimated by the following formula: $\hat{\pi}(\hat{N}_i) = \pi(\hat{\mu}, \hat{N}_i) = e^{\hat{\mu}+\hat{N}_i}/(1 + e^{\hat{\mu}+\hat{N}_i})$, where the estimated values for $N_i$ and $\mu$ are listed in Table 5.

**Table 5. Estimated node effects in the logistic regression model using recombinant inbred lines.**

| Node | Estimate | |
|---|---|---|
| Intercept | −1.76 | ** |
| 5 | 0.33 | ** |
| 6 | 1.21 | ** |
| 7 | 1.46 | ** |
| 8 | 1.43 | ** |
| 9 | 1.33 | ** |
| 10 | 1.27 | ** |
| 11 | 1.25 | ** |
| 12 | 1.09 | ** |
| 13 | 0.99 | ** |
| 14 | 0.83 | ** |
| 15 | 0.68 | ** |
| 16 | 0.47 | ** |
| 17 | −0.07 | * |
| 18 | −0.67 | ** |
| 19 | −1.51 | ** |
| 20 | −2.28 | ** |
| 21 | −3.56 | ** |

* Significant at 0.05 level of probability.
** Significant at 0.01 level of probability.

## Comparisons between Mixed Linear Model and Logistic Regression Model

Mean boll retention and standard errors for different nodes with RI population were estimated for mixed models and logistic regression models. Estimated boll retention appeared to be very similar for the two models (Table 6), suggesting that both mixed models and logis-

**Table 6. Estimated first-position boll retention and their 95% confidence interval length (CIL) on different nodes.**

| | CSH† | | ARH(1)‡ | | UN§ | | CS¶ | | Logistic | |
|---|---|---|---|---|---|---|---|---|---|---|
| Node | EST | CIL | EST | CIL | EST | CIL | EST | CIL | EST | CIL |
| | % | | | | | | | | | |
| 5 | 20.14 | 2.95 | 20.14 | 2.79 | 20.14 | 2.81 | 20.14 | 2.25 | 19.36 | 1.71 |
| 6 | 37.58 | 3.21 | 37.58 | 3.13 | 37.58 | 3.15 | 37.58 | 2.25 | 36.52 | 2.08 |
| 7 | 43.40 | 2.75 | 43.40 | 2.71 | 43.40 | 2.73 | 43.40 | 2.25 | 42.61 | 2.14 |
| 8 | 41.98 | 2.89 | 41.98 | 2.87 | 41.98 | 2.88 | 41.98 | 2.25 | 41.83 | 2.14 |
| 9 | 39.55 | 2.87 | 39.55 | 2.95 | 39.55 | 2.83 | 39.55 | 2.25 | 39.44 | 2.12 |
| 10 | 37.77 | 2.56 | 37.77 | 2.76 | 37.77 | 2.55 | 37.77 | 2.25 | 38.07 | 2.10 |
| 11 | 37.60 | 2.61 | 37.60 | 2.74 | 37.60 | 2.56 | 37.60 | 2.25 | 37.54 | 2.10 |
| 12 | 33.60 | 2.39 | 33.60 | 2.62 | 33.60 | 2.39 | 33.60 | 2.25 | 33.82 | 2.05 |
| 13 | 31.40 | 2.36 | 31.40 | 2.62 | 31.40 | 2.35 | 31.40 | 2.25 | 31.63 | 2.01 |
| 14 | 27.63 | 2.28 | 27.63 | 2.46 | 27.63 | 2.29 | 27.63 | 2.25 | 28.21 | 1.95 |
| 15 | 24.74 | 2.20 | 24.74 | 2.27 | 24.74 | 2.23 | 24.74 | 2.25 | 25.27 | 1.88 |
| 16 | 21.28 | 2.44 | 21.28 | 2.35 | 21.28 | 2.48 | 21.28 | 2.25 | 21.61 | 1.78 |
| 17 | 13.55 | 2.10 | 13.55 | 1.96 | 13.55 | 2.14 | 13.55 | 2.25 | 13.80 | 1.49 |
| 18 | 8.13 | 1.73 | 8.13 | 1.59 | 8.13 | 1.74 | 8.13 | 2.25 | 8.13 | 1.18 |
| 19 | 3.65 | 0.93 | 3.65 | 0.89 | 3.65 | 0.95 | 3.65 | 2.25 | 3.65 | 0.81 |
| 20 | 1.81 | 0.65 | 1.81 | 0.64 | 1.81 | 0.67 | 1.81 | 2.25 | 1.73 | 0.57 |
| 21 | 0.46 | 0.33 | 0.46 | 0.33 | 0.46 | 0.34 | 0.46 | 2.25 | 0.49 | 0.31 |
| 22 | 0.24 | 0.22 | 0.24 | 0.22 | 0.24 | 0.22 | 0.24 | 2.25 | 0.24 | 0.22 |

† CSH = heterogeneous compound symmetry.
‡ ARH(1) = heterogeneous autoregressive(1).
§ UN = unstructured.
¶ CS = compound symmetry.

tic model could offer similar estimates. Mean 95% CIL was 2.08, 2.11, 2.07, 2.25, and 1.59% for the CSH, ARH(1), UN, CS, and logistic, respectively. On the other hand, the logistic regression model gave smaller CIL than the mixed linear model with four types of error structures for each estimate of boll retention (Table 6). It suggested that the use of some error structures in the mixed linear model may provide a higher precision than the use of a linear model approach.

## DISCUSSION

Understanding the properties of space-dependent boll retention is useful for breeders and growers to understand breeding and production. In previous studies, this trait was considered as a continuously and normally distributed variable. In this study, we looked at boll retention differently and analyzed it using a logistic regression model. This study showed that the logistic regression model gave similar mean estimates as the mixed model with different error structures but provided different confidence intervals. Although we found that the logistic regression model provided the smaller confidence intervals, we still cannot prove that the logistic model is better than the other. The reasons are possibly that the mixed linear approach (including ANOVA method) calculates the standard error of each least square mean based on the residual variance while the standard error for an estimated probability in a logistic regression model depends on the estimated probability and the number of plants observed.

Boll retention for the first position in the middle of the plant for the RI lines in this study was lower than that for widely grown cultivars (Jenkins and McCarty, 1995). Retention could be improved through a proper breeding scheme and/or improvement of environmental conditions. Due to the instability, small-sized bolls, and poor fiber quality for Node 16 and above in cotton production, bolls produced from these nodes usually can be ignored. High boll retention for Nodes 6 through 15 should be very important for improving cotton production because bolls in the middle of a plant normally yield better fiber and account for the major contribution to total cotton yield (Jenkins and McCarty, 1995). Little difference was found for boll retention at the first position for Nodes 7 to 14 among commercial cultivars except for the short-season genotypes DH 126 and DES119 and full-season cultivars DP90 and DP5690 (Jenkins and McCarty, 1995). Similar results were found in this study; however, the unpublished data and many other studies showed that total boll number and cotton yield were mainly controlled by genotypic effects. Possibly, bolls (or boll retentions) at the second position may make yield differences among genotypes. Thus, the potential for boll retention at other positions, such as second position, on the middle nodes of a plant could be an important consideration for yield improvement. Field management should focus on how to improve boll retention probability on the first two positions.

The main objective was to compare the results between the mixed linear model and the logistic regression model. Due to the time and labor required to collect boll retention data for 191 cotton lines, this investigation was conducted only for the first position in 1 yr; however, we believe that the results obtained from a large data set provided reliable information for comparing the two statistical models to evaluate boll retention and possibly other binary traits. If genotype × environment interactions have strong impacts on boll retention, repeating the experiment in multiple environments would be needed.

## REFERENCES

Agresti, A. 1990. Categorical data analysis. John Wiley & Sons, New York.

Agresti, A. 1996. An introduction to categorical data analysis. John Wiley & Sons, New York.

Collett, D. 1991. Modelling binary data. Chapman and Hall, London, UK.

Cox, D.R., and E.J. Snell. 1989. The analysis of binary data. 2nd ed. Chapman and Hall, London, UK.

Fehr, W.R. 1987. Principles of cultivar development: Volume 1. Theory and technique. Macmillan Publ. Co., New York.

Hosmer, D.W., Jr., and S. Lemeshow. 1996. Applied logistic regression. John Wiley & Sons, New York.

Jenkins, J.N., and J.C. McCarty, Jr. 1995. End of season plant maps. Bull. 1024. Mississippi Agric. & Forestry Exp. Stn., Mississippi State.

Jenkins, J.N., J.C. McCarty, Jr., and W.L. Parrot. 1990a. Effectiveness of fruiting sites in cotton: yield. Crop Sci. 30:365–369.

Jenkins, J.N., J.C. McCarty, Jr., and W.L. Parrot. 1990b. Fruiting efficiency in cotton: Boll size and boll set percentage. Crop Sci. 30: 857–860.

Kerby, T.A., J. Keeley, and S. Johnson. 1987. Growth and development of *acala* cotton. Bull. 1921. Univ. of California Agric. Exp. Stn. Div. of Agric. and Nat. Res., Oakland.

Littell, R.C., G.A. Milliken, W.W. Stroup, and R.D. Wolfinger. 1996. SAS system for mixed models. SAS Inst., Cary, NC.

McCullagh, P., and J.A. Nelder. 1989. Generalized linear models. Chapman Hall, London, UK.

SAS Institute Inc. 1999a. SAS OnLine Doc version 8.0. SAS Inst., Cary, NC.

SAS Institute Inc. 1999b. SAS software version 8.0. SAS Inst., Cary, NC.

Shappley, Z.W., J.N. Jenkins, W.R. Meredith, and J.C. McCarty, Jr. 1998a. An RFLP linkage map of upland cotton (*Gossypium hirsutum* L.). Theor. Appl. Genet. 97:756–761.

Shappley, Z.W., J.N. Jenkins, J. Zhu, and J.C. McCarty, Jr. 1998b. Quantitative traits loci associated with agronomic and fiber traits of upland cotton. J. Cotton Sci. 4:153–163.

Shoemaker, D. 2000. Genetic analysis of agronomic traits of selected American and Australian cotton genotypes and their $F_2$ hybrids. Ph.D. diss., Mississippi State Univ., Mississippi State.